Title: "Theory, practice, and conservation in the age of genomics: the Galápagos giant tortoise as a case study"

Authors:

Stephen J Gaughran[1]*, Maud C Quinzin[1], Joshua M Miller[1], Ryan C Garrick[2], Danielle L Edwards[3], Michael A Russello[4], Nikos Poulakakis[5,6], Claudio Ciofi[7], Luciano B Beheregaray[8], Adalgisa Caccone[1]*

1. Department of Ecology and Evolutionary Biology, Yale University, 21 Sachem St. New Haven, Connecticut, 06520, United States of America
2. Department of Biology, University of Mississippi, Oxford, Mississippi, 38677, United States of America
3. Life and Environmental Sciences, University of California, Merced, 5200 N Lake Rd, Merced, California, 95343, United States of America
4. Department of Biology, University of British Columbia, Okanagan Campus, Kelowna, BC V1V 1V7, Canada
5. Department of Biology, School of Sciences and Engineering, University of Crete, Vasilika Vouton, Gr-71300, Heraklio, Crete, Greece
6. Natural History Museum of Crete, School of Sciences and Engineering, University of Crete, Knossos Av., GR-71409, Heraklio, Crete, Greece
7. Department of Biology, University of Florence, 50019 Sesto Fiorentino (FI), Italy
8. Molecular Ecology Lab, School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA, 5001, Australia

*Corresponding authors: stephen.gaughran@yale.edu; adalgisa.caccone@yale.edu

Abstract:

High-throughput DNA sequencing allows efficient discovery of thousands of single nucleotide polymorphisms (SNPs) in non-model species. Population genetic theory predicts that this large number of independent markers should provide detailed insights into population structure, even when only a few individuals are sampled. Still, sampling design can have a strong impact on such inferences. Here, we use simulations and empirical SNP data to investigate the impacts of sampling design on estimating genetic differentiation among populations that represent three species of Galápagos giant tortoises (*Chelonoidis* spp.). Though microsatellite and mitochondrial DNA analyses have supported the distinctiveness of these species, a recent study called into question how well these markers matched with data from genomic SNPs, thereby questioning decades of studies in non-model organisms. Using >20,000 genome-wide SNPs from 30 individuals from three Galápagos giant tortoise species, we find distinct structure that matches the relationships described by the traditional genetic markers. Furthermore, we confirm that accurate estimates of genetic differentiation in highly structured natural populations can be obtained using thousands of SNPs and 2-5 individuals, or hundreds of SNPs and 10 individuals, but only if the units of analysis are delineated in a way that is consistent with evolutionary history. We show that the lack of structure in the recent SNP-based study was likely due to unnatural grouping of individuals and erroneous genotype filtering. Our study demonstrates that genomic data enable patterns of genetic differentiation among populations to be elucidated even with few samples per population, and underscores the importance of sampling design. These results have specific implications for studies of population structure in endangered species and subsequent management decisions.

"Modern molecular techniques provide unprecedented power to understand genetic variation in natural populations. Nevertheless, application of this information requires sound understanding of population genetics theory."

- Fred Allendorf (2017, p. 420)

Introduction:

The advent of high-throughput DNA sequencing has enabled the characterization of the genomes of model and non-model organisms alike. Genome-wide data can improve the precision and accuracy of estimates of population parameters, enhancing our understanding of present-day structure, gene flow, and local adaptation (Funk *et al.* 2012). These data have also facilitated more detailed reconstructions of historical events that impacted evolutionary trajectories within species (e.g., Emerson et al. 2010), and among closely related species (e.g., Chaves et al. 2016).

While whole genome sequencing is still beyond the budget of many research programs, methods based on reduced representation genomic libraries (e.g., double digest Restriction-site Associated DNA sequencing, ddRADseq (Peterson *et al.* 2012)) allow tens or hundreds of thousands of single nucleotide polymorphisms (SNPs) to be discovered and reliably genotyped at a much-reduced cost (Andrews *et al.* 2016). This is particularly beneficial for species of conservation concern, where limited resources and sampling constraints (i.e., few individuals are available) may be prevalent. No matter the application, though, well-designed population genetics studies aim to maximize their statistical power while minimizing costs.

Genome-wide SNP data are currently being applied to a broad spectrum of conservation objectives. These range from informing captive breeding programs (e.g., Wright et al. 2015) and improving detection of hybridization and inbreeding depression (e.g., Robinson et al. 2016; vonHoldt et al. 2016b), to delineating conservation units, assessing levels of adaptive genetic

variation, and predicting viability in the face of anthropogenic impacts such as climate change (Henry & Russello 2013; Rellstab *et al.* 2015; Sork *et al.* 2016; Brauer *et al.* 2016). The appeal of genomic approaches to conservation biology is heightened by indications that a large number of independent loci can alleviate issues associated with small sample sizes per population; when using thousands of loci one can obtain reliable estimates of genetic diversity and population differentiation, so long as the true values of these parameters are sufficiently high (e.g., Li and Durbin 2011; Willing et al. 2012). Yet, as noted by Allendorf (2017), genomic datasets need to be analyzed within the context of a carefully considered sampling design. Shortcomings in sampling design can lead to erroneous conclusions (Meirmans 2015), which can have profound consequences for any population level study, but especially for those with direct management implications for threatened or endangered species.

Here, we explore the power of using thousands of SNP markers to study population structure, and the impact of sampling design and small sample sizes on detecting and describing that structure. To do this, we use genomic data from Galápagos giant tortoises (*Chelonoidis* spp.) as a case study, given a recent study has questioned the genomic distinctiveness of several species within this genus (Loire *et al.* 2013). The Galápagos Islands are home to a radiation of endemic giant tortoises that includes 11 endangered and 4 extinct species (Fig. 1). Taxonomic designations are supported by differences in morphology, geographic isolation of most species, and evidence of evolutionary divergence based on mitochondrial DNA (mtDNA) and nuclear microsatellite data ((Ciofi et al. 2002; Beheregaray et al. 2003a; Garrick et al. 2015); see Fig. S7 A and B).

In contrast to previous studies (see supplementary material section VIII for details; Ciofi et al. 2002; Beheregaray et al. 2003a; Beheregaray et al. 2004; Russello et al. 2005; Poulakakis et al. 2012; Garrick et al. 2015; Poulakakis et al. 2015), Loire et al. (2013) challenged the genetic distinctiveness of three Galápagos giant tortoise species. Those authors collected

100 transcriptome-derived genotypic data from ~1000 synonymous SNPs from five captive
101 individuals representing three species (*C. becki*, *C. porteri* and *C. vandenburghi*). They did not
102 detect significant differentiation, as measured by F_{ST}, when comparing two groups (one group of
103 three *C. becki* individuals, and the second group consisting of two individuals, one *C. porteri* and
104 one *C. vandenburghi*). These two groups were constructed on what the authors identified as
105 natural partitions, based on the observation that their samples fall into two different mtDNA
106 clades (Fig. S6a; Poulakakis et al. 2012). Furthermore, Loire et al. (2013) did not detect
107 homozygosity excess, as measured by F_{IT}, for which positive values would indicate population
108 structure. Given that previous population genetic studies have largely relied upon data from
109 mtDNA and microsatellites, such a discrepancy between these traditional markers and genomic
110 SNPs could have wide-ranging implications, beyond the case of Galápagos giant tortoises, and
111 therefore warrants further investigation.

112 In this study, we investigate the agreement of population structure analyses based on
113 genome-wide SNPs compared to those based on mtDNA sequences and microsatellite
114 genotypes. To do this we generated a dataset of tens of thousands of genome-wide SNPs from
115 30 individuals representing the same three species (*C. becki*, *C. porteri*, and *C. vandenburghi*)
116 considered by Loire et al. (2013). Since these species form a recently diverged species complex,
117 we treat each species as a population to compare against the null hypothesis that all Galápagos
118 giant tortoises belong to a single species with one panmictic population. First, we address
119 whether or not there is significant genomic differentiation among these three Galápagos giant
120 tortoise species using newly generated SNPs. Then, we subsample our data to explore the
121 effects of using only a few individuals per population and of pooling individuals from different
122 populations on estimating genetic differentiation. From these subsampling simulations, we
123 predict the range of F_{ST} estimates expected when using the sampling scheme of Loire et al.

(2013). Finally, we reanalyze the raw RNA-seq data from Loire et al. (2013) to test our prediction.

Materials and Methods:

Sampling and sequencing

Samples were obtained during previously conducted collection expeditions (Caccone et al. 1999; Caccone et al. 2002; Ciofi et al. 2002; Beheregaray et al. 2003a; Beheregaray et al. 2003b; Beheregaray et al. 2004; Russello et al. 2005; Ciofi et al. 2006; Russello et al. 2007; Poulakakis et al. 2008; Garrick et al. 2012; Poulakakis et al. 2012; Edwards et al. 2013; Edwards et al. 2014; Garrick et al. 2014). Approximately ten samples per population for each extant species (n=121 individuals in total) were selected for sequencing as part of a larger project on the phylogeography of Galápagos giant tortoises. These individuals were chosen as they displayed concordant and unambiguous genetic assignments between mitochondrial (control region, mt*CR*) and microsatellite (12 loci) ancestry based on a published database of 123 mitochondrial haplotypes (Poulakakis *et al.* 2012) and 305 genotyped individuals (Edwards *et al.* 2013) that include all the extant and extinct populations and species.

DNA was extracted from blood samples using a DNeasy Blood and Tissue kit (Qiagen) according to the manufacturer's instructions. We then prepared ddRAD libraries following Peterson et al. (2012). For each sample, 500 ng of genomic DNA was digested with the restriction enzymes *MluCI* and *NlaIII* (New England BioLabs), and ligated with Illumina-specific adaptors representing up to 18 unique barcodes and 2 index codes. Ligated fragments of samples were pooled into 13 libraries and size-selected to be ~310 bp (range 279 – 341bp) with a BluePippin (Sage Science). Size-selected libraries included 12 to 24 individuals and were

148 paired-end sequenced on 13 lanes of an Illumina HiSeq 2000 at the Yale Center for Genome
149 Analysis.
150
151 SNP calling
152 We used forward and reverse reads to generate a *de novo* assembly using the pyrad v.3.0.3
153 pipeline (Eaton 2014). Reads were de-multiplexed and assigned to each individual based on
154 barcodes allowing for one mismatch. We replaced base calls of Q < 20 with an ambiguous base
155 (N) and discarded sequences containing more than four ambiguities. We used 85% clustering
156 similarity as a threshold to align the reads into loci. We set additional filtering parameters to
157 allow for a maximum number of SNPs to be called: retaining clusters with a minimum depth of
158 sequence coverage (Mindepth) > 5 and a locus coverage (MinCov) > 10, a maximum proportion
159 of individuals with shared heterozygote sites of 20% (MaxSH = p.20), and a maximum number
160 of SNP per locus of 15 (maxSNP = 15). For subsequent analyses, we filtered this dataset using
161 vcftools (Danecek *et al.* 2011) to generate a set of polymorphic loci (23,057 SNPs) with no
162 missing data common to all three Galápagos giant tortoises populations of interest, abbreviated
163 PBL, CRU, and VA and corresponding to the species *C. becki*, *C. porteri*, and *C. vandenburghi*,
164 respectively (n = 10 individuals each).
165
166 Analytical methods
167 *F*-statistics (F_{IT}, F_{IS}, global F_{ST}, and pairwise F_{ST}) were calculated using the diveRsity
168 package in R (Keenan *et al.* 2013), which uses a weighted Weir and Cockerham (1984)
169 estimator. The same package was used to assess the statistical significance of these estimates
170 by bootstrapping across loci. Through this method we established 95% confidence intervals for
171 each estimate, accepting as significant those that did not include 0. Pairwise F_{ST} calculated from
172 thousands of subsamples of the data (described below) were carried out in vcftools (Danecek *et*

al. 2011) to streamline computation. We also used vcftools to calculate the number of loci out of Hardy-Weinberg equilibrium for each population and for pooled populations.

Since F_{ST} estimates rely on *a priori* assignment of individuals to groups that are typically based on geographic location, we used two methods that do not have this assumption to assess patterns of differentiation among our samples. To do this, we first carried out Principal Component Analysis (PCA) on all 30 individuals, using the PLINK software (Chang *et al.* 2015). Principal components 1 and 2 were plotted against each other in R. To complement the multivariate analyses, we performed a Bayesian clustering analysis, implemented in the program STRUCTURE version 2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003), also including all 30 individuals. STRUCTURE assumes a model with K unknown clusters representing genetic populations in Hardy-Weinberg equilibrium, and then assigns individuals to each cluster based on allele frequencies. We ran 20 repetitions of STRUCTURE for K=1–5, with a burn-in of 10,000 iterations and MCMC length of 50,000 iterations. These runs used the admixture model, correlated allele frequencies among populations, and did not assume prior population information. All other parameters were left at default values. Results were post-processed and visualized using CLUMPAK (Kopelman *et al.* 2015). We used mean log likelihood values (Pritchard *et al.* 2000) and the ΔK statistic (Evanno *et al.* 2005) to infer the best K (Supplementary Figure S5). Both analyses considered the 23,057 SNPs common to all individuals.

To further assess the power of our SNPs to detect population structure, we randomly subsampled individuals from each of the species and calculated pairwise F_{ST} for each species using these subsamples. We tested this for per-species sample sizes of n=2, n=3, and n=5. This process was repeated 1,000 times for each sample size. We also carried out a similar analysis maintaining all 10 individuals per population but randomly subsampling SNPs from our dataset. For these analyses we used the following number of SNPs: 25, 50, 100, 200, 500, 1000,

198 5000, and 10,000. This was repeated 1,000 times for each sample size. Finally, we used a
199 subsampling scenario that directly mimicked the one in Loire *et al.* (2013) to further evaluate the
200 impact of limited sample sizes and pooling of samples from distinct species on F_{ST} estimates.
201 As was done in Loire *et al.* (2013), we compared a set of three individuals from *C. becki* to a
202 grouping that included one *C. porteri* plus one *C. vandenburghi* individual. To account for
203 sample variation, we repeated this grouping process 1,000 times (described in full in
204 supplementary material section IV).
205
206 **Results:**
207 Tortoise samples and ddRAD-seq dataset
208 Our sequencing generated a total of 3,094,399,092 retained reads (approximately 15 to
209 58 million reads per individual) after de-multiplexing and filtering reads for quality and
210 ambiguous barcodes and ddRAD-tags. *de novo* assembly of the data resulted in 48,004,056
211 ddRAD-tags (approximately 320,000-465,000 per individual). From these, we called SNPs and
212 obtained 973,321 variable sites. We then narrowed those loci down to only loci with genotypes
213 called in every individual in our three species data set, for a total of 23,057 SNPs. For the three
214 species of interest the number of loci retained within populations and between populations pairs
215 are presented in Table 1. The average coverage per locus per individual was 12X (minimum 9;
216 maximum 15).
217
218 *F*-statistics using ddRAD-seq data
219 Calculation of *F*-statistics revealed values consistent with highly-structured populations
220 (F_{IT} = 0.257, 95% CI: 0.251 – 0.262; F_{IS} = 0.079, 95% CI: 0.073 – 0.084; and global F_{ST} = 0.193,
221 95% CI: 0.189 – 0.198). Using the SNPs in common to each population pair (Table 1), we found
222 pairwise F_{ST} values of 0.169 (95% CI: 0.164 – 0.174) between PBL and CRU, 0.181 (95% CI:

223 0.175 – 0.187) between PBL and VA, and 0.233 (95% CI: 0.226 – 0.240) between CRU and VA
224 (Table 2). These estimates were similar to, though higher than, F_{ST} estimates using 12 nuclear
225 microsatellite markers (Garrick *et al.* 2015) for these species comparisons (Table 2 and
226 supplementary materials Table S5).
227
228 PCA and STRUCTURE
229 The first two principal components of the PCA showed clear differentiation among
230 individuals from the three species. PC1 accounted for approximately 12.0% of the variation
231 among individuals and PC2 accounted for approximately 9.3% of the variation among
232 individuals (Figure 2). Similarly, both mean log likelihood values (Pritchard *et al.* 2000) and the
233 ΔK statistic (Evanno *et al.* 2005) supported the existence of three distinct genetic units in the
234 STRUCTURE analysis (Supplementary Figure S5). These groups correspond to the *a priori*
235 geographic groupings used in F_{ST} estimates and to the three named species. Our separate
236 analysis of loci out of Hardy-Weinberg equilibrium (HWE), the basis for the STRUCTURE
237 algorithm, supported these findings as well. When each species was considered separately, out
238 of 23,057 loci PBL showed 214 out of HWE, CRU showed 124 out of HWE, and VA showed 71
239 out of HWE. When the CRU and VA samples were pooled, the number of loci out of HWE rose
240 to 1326. When all three species were pooled and treated as one population, 2422 loci were
241 found to be out of HWE.
242
243 Sample size, number of loci, and the effect of individual samples
244 In all population comparisons for the three sample sizes (n = 2, 3, or 5), the majority of
245 estimates were within 0.03 of the F_{ST} value calculated using the complete dataset of 10 samples
246 per population (Figure 3). In every case, when the sample size was two F_{ST} tended to be
247 underestimated, though with a long tail of overestimated outliers. In all comparisons with sample

248 sizes of three or five this skew disappeared: we found that 95% of the estimates were within
249 0.05 of the estimate using 10 samples (Supplementary Tables S1).

250 Our F_{ST} estimates from subsampled SNPs ranging from 25 to 10,000 SNPs appeared to
251 have the statistical power to detect population structure between these population pairs when
252 10 individuals were used, with 95% of all estimates above 0 (Table S2 A – C). However, as
253 expected, with many fewer SNPs the range of 95% of the estimates was very wide (see
254 supplementary material section III). For example, when only 100 SNPs were used to compare
255 PBL and CRU, 95% of the F_{ST} estimates were between 0.1 and 0.255, while using 1000 SNPs
256 gave 95% of the F_{ST} estimates between 0.146 and 0.194 for the same comparison (Table S2A).
257
258 Effect of pooling samples
259 To test how pooling samples affected F-statistic estimates, we used the Loire *et al.*
260 (2013) sampling design, pooling one individual from *C. porteri* and one individual from *C.*
261 *vandenburghi* into one population and comparing this to three individuals from *C. becki*. When
262 the set of common SNPs (*n*=23,057) were included in the analysis, the F_{ST} estimates between
263 1000 pairs of these groups ranged from 0.045 to 0.136 (95%: 0.052–0.127, mean: 0.075). When
264 only 1,000 SNP loci were used, as in Loire *et al.* (2013), the F_{ST} estimates ranged from 0.006 to
265 0.157 (95%: 0.031–0.134, mean: 0.076) (see Supplementary Figures S4A and S4B). This
266 confirms that pooling samples from two populations, each representing different species, results
267 in a strongly depressed F_{ST} estimate. However, these simulations highlight that the occurrence
268 of genetic differentiation (i.e., positive F_{ST} values) should still be detectable even with this
269 grouping scheme.
270
271 Re-analysis of Loire et al. transcriptome data

Given that our analyses of ddRAD-seq data showed clear genetic structure among the populations from the three species, and our subsampling simulations (Figure S4) predicted that positive F_{ST} values should still be detectable using the grouping scheme adopted by Loire et al. (2013), we reanalyzed the original RNA-seq data generated for that publication to further assess the source of the discrepancy. We downloaded the publically-available RNA sequencing data generated by Loire et al. (2013) from the NCBI's Sequence Read Archive and re-called SNPs after aligning these reads to a draft genome assembly of a closely related species of Galápagos giant tortoise, *C. abingdonii* (unpublished data; see methods in supplementary materials section VII). With these transcriptome-derived SNP data, we estimated an F_{ST} of 0.054 (95% CI: 0.049 – 0.058) when comparing the three *C. becki* samples (PBL) to the combined two *C. porteri* and *C. vandenburghi* samples (CRU and VA). Notably, this F_{ST} value falls within our predicted range of F_{ST} estimates generated by subsampling the ddRAD-seq data. Our F_{IT} estimate for this data set was -0.121 (95% CI: -0.129 – -0.113), with F_{IS} estimated to be -0.185 (95% CI: -0.192 – 0.177).

Plotting the first two principal components of a PCA of these five samples showed clear clustering of the conspecific samples from *C. becki*, while the single samples from *C. vandenburghi* (VA) and *C. porteri* (CRU) are distinct from each other and from the *C. becki* samples (supplementary Figure S6).

Discussion

Strong evidence of population structure

Using genome-wide SNP data we found evidence for significant differentiation among the three species considered (*C. becki*, *C. porteri*, and *C. vandenburghi*), consistent with the findings of decades of research in this system (Ciofi et al. 2002; Beheregaray et al. 2003a; Beheregaray et al. 2003b; Beheregaray et al. 2004; Russello et al. 2005; Russello et al. 2007;

297 Poulakakis et al. 2008; Poulakakis et al. 2012; Garrick et al. 2015; Poulakakis et al. 2015). Our
298 estimate of F_{IT} (0.257), which was a focal metric used in the previous study (Loire *et al.* 2013),
299 was positive and significantly different from zero. Positive values of F_{IT} indicate an excess of
300 homozygous loci in the sample set. This could suggest the existence of population structure in
301 the total sample set. This possibility is reinforced by the finding of very high and
302 significantly different from zero F_{ST} estimates for the same comparisons (between 0.17
303 and 0.24; Table 2, supplementary Figure S1). Interpreting significantly positive F_{IS} values,
304 such as the one calculated from our ddRAD-seq data set, can be difficult (Allendorf and Luikart
305 2007). This could be due to substructure within one or more populations, sampling stochasticity,
306 and/or recent demographic changes in relatively small populations. It could also be that such
307 small populations are not necessarily expected to be in Hardy-Weinberg equilibrium due to the
308 increased influence of genetic drift (Allendorf & Luikart 2009).

309 To assess whether there was additional genetic structure outside of our *a priori*
310 assignment of individuals based on their geographic location, we also analyzed the 30 samples
311 in our ddRAD-seq data set using two methods without prior assignment of each sample to a
312 group. Both principal component (Figure 2) and Bayesian clustering analyses (supplementary
313 Figure S5) clearly discerned three genetically distinct clusters that corresponded to the samples
314 from the three species tested in our pairwise F_{ST} estimates. This echoed our per-locus analysis
315 of Hardy-Weinberg equilibrium (HWE), which showed that treating all 30 individuals from the
316 three named species as a single population dramatically increased the number of loci out of
317 HWE.

318 Results of our analyses of population structure using tens of thousands of genome-wide
319 SNPs are concordant with earlier studies using mtDNA haplotypes and microsatellite genotypes
320 (Ciofi et al. 2002; Beheregaray et al. 2003a; Beheregaray et al. 2003b; Beheregaray et al. 2004;

Russello et al. 2005; Russello et al. 2007; Poulakakis et al. 2008; Poulakakis et al. 2012; Garrick et al. 2015; Poulakakis et al. 2015). These findings definitively resolve concerns raised by Loire et al. (2013) regarding whether these traditional markers were accurately reflecting the genetic distinctiveness of Galápagos giant tortoise species. Importantly, our results not only revealed the same genetic clustering as earlier studies, but also showed the same patterns of genetic distance. As in the microsatellite studies, we found slightly greater genetic differentiation between *C. becki* and *C. vandenburghi* (PBL and VA: F_{ST} = 0.181) than between *C. becki* and *C. porteri* (PBL and CRU: F_{ST} = 0.169), and the greatest differentiation between *C. porteri* and *C. vandenburghi* (CRU and VA: F_{ST} = 0.233) (Table 2). While qualitatively the same, our F_{ST} estimates are notably higher than those calculated using microsatellites (Table 2), a finding predicted by the mathematics of using biallelic vs. multiallelic loci (Putman & Carbone 2014), which has also been found in other systems (e.g., Payseur and Jing 2009).

Impact of sample size and number of loci on detecting population structure

Population genetic theory (Nei 1978), simulations (Willing *et al.* 2012), and empirical work (Reich *et al.* 2009) support the idea that a data set of thousands of loci should have the power to detect population structure with high precision, even when only a few individuals per population are analyzed. We tested this idea with our Galápagos giant tortoise ddRAD-seq SNP data by estimating F_{ST} from subsamples of two, three, and five individuals from each population and comparing them to the same estimates obtained from 10 individuals per population. All tested sample sizes were able to detect significant F_{ST} values, though using three or five samples yielded more precise estimates than using only two (Figure 3; supplementary tables S1). These analyses are consistent with the idea that accurate F_{ST} values can be estimated using as few as two or three samples per population if thousands of SNPs are analyzed. Likewise, we found that for highly differentiated populations such as those studied here,

346 hundreds of SNPs were sufficient to accurately describe population structure when ten
347 individuals per population were used. This empirical evidence should be helpful in the design of
348 future conservation genetics studies that aim to describe population structure, in which case
349 additional samples may lead to diminishing returns for improving statistical power. This will be
350 especially useful for endangered or elusive species for which sampling may present a severe
351 limitation.
352
353 Sampling design matters
354 Our genome-wide SNP data detected high and significant differentiation among these
355 three species, even when only two or three individuals from each were used in the analysis
356 (Figure 3). While these results were strongly supported, they failed to explain the discrepancy
357 described by Loire et al. (2013), who used over 1,000 synonymous SNPs from transcriptome
358 sequencing data and found no differentiation between the same three species. Their sample
359 size of five captive individuals does not by itself account for the discrepancy between the two
360 studies, because, as we show above (supplementary Figure S4), using thousands of SNPs
361 should give sufficient power to detect population structure in Galápagos giant tortoises, even
362 when sample size is that small.
363 Instead, sampling design, and specifically grouping of individuals into inappropriate
364 population units, rather than sample size likely biased the statistical power of Loire et al.'s
365 (2013) study. Their sampling scheme divided the five individuals into two groups, which did not
366 reflect the population divergence of the three species. Specifically, this mixed group included
367 two individuals, each from different species (CRU, *C. porteri* from Santa Cruz Island; VA, *C.*
368 *vandenburghi* from central Isabela Island), and another group of three individuals from the other
369 species (PBL, *C. becki* from northern Isabela Island). The justification for this grouping was
370 based on the closer phylogenetic relationship of mtDNA haplotypes from *C. porteri* and *C.*

371 *vandenburghi* (Caccone et al. 1999; Russello et al. 2007) compared to haplotypes found in the
372 PBL *C. becki* population. This choice is problematic for several reasons (detailed in the
373 supplementary material section VIII). Most importantly, *F*-statistics are a reflection of population
374 differentiation, not of phylogenetic relatedness. Treating the individuals from *C. porteri* and *C.*
375 *vandenburghi* as belonging to the same population biased the *F*-statistics estimates by leading
376 to an increase in within-group variation, and therefore depressed F_{ST} values. This within-group
377 structure, which distorts *F*-statistics, is known as Wahlund effect (Wahlund 1928).

378 The problem outlined above is clear in our pairwise analysis using >20,000 SNPs, which
379 shows that while the *C. becki* population sample is about equally differentiated from the *C.*
380 *porteri* and *C. vandenburghi* ones, the ones from *C. porteri* and *C. vandenburghi* are more
381 differentiated from each other than from the *C. becki* population sample (Table 2). To empirically
382 test for the Wahlund effect under this sampling scheme, we simulated a scenario in which three
383 samples from *C. becki* were compared to a population consisting of one *C. porteri* and one *C.*
384 *vandenburghi* sample. Repeating this sampling scenario 1,000 times, we found significantly
385 depressed mean F_{ST} estimates, as low as 0.075, with 95% of comparisons ranging from 0.052
386 to 0.127 (supplementary Figure S4A). Even more strikingly, when we limited the analysis to a
387 similar number of markers as Loire et al. (2013) and used 1,000 randomly drawn SNPs, the
388 range of 95% of the estimates increased to 0.031 to 0.134.

389

390 <u>RNA-seq data supports population structure</u>

391 While our subsampling simulations showed a clear Wahlund effect when samples from
392 two different species (*C. porteri* and *C. vandenburghi*) were combined into one grouping, these
393 F_{ST} estimates were still positive (mean F_{ST} = 0.075). We therefore would have expected Loire et
394 al. (2013) to find a similar estimate in their analysis of RNA-seq data, but they reported no
395 significantly positive F_{ST} value. To investigate this discrepancy, we re-analyzed their raw

sequencing data by aligning it to a Galápagos giant tortoise reference genome. Using the SNPs from this reanalysis, we estimated an F_{ST} of 0.054, which is similar to our expected F_{ST} under their sampling design (supplementary Figure S4). Our estimates of F_{IS} and F_{IT} for the RNA-seq data set were negative, a surprising result that may be related to the sampling design, the specific individuals included in that study, or the deviations from Hardy-Weinberg equilibrium that can occur in small populations (Kimura & Crow 1963). This last point is due to the assumption of large numbers in Hardy-Weinberg equilibrium, which is violated in small populations (Allendorf & Luikart 2009).

Convincingly, a PCA of Loire et al.'s (2013) SNP data revealed a tight cluster of the three PBL samples, whereas the CRU and VA samples were distinct both from the PBL cluster and from each other (Figure S6). This pattern of principal components mirrors the one that we found with our 30 sample dataset for the same populations (Figure 2). These results, which match our expectations based on subsampling simulations (Supplementary figure S4), suggest that the lack of significantly positive F_{ST} values found by Loire et al. (2013) is due not just to small sample size and inappropriate grouping of samples, but also the genotype filters employed in their initial analysis. The original Loire et al. (2013) methods describe a genotype filter that assigns posterior probability to genotypes based on Hardy-Weinberg equilibrium. We suspect that this may not be a reliable method when genotyping a pool of individuals from different species, since these samples will not meet the assumption of Hardy-Weinberg equilibrium. Our SNP calls of their data may have also been improved by mapping the RNA sequence reads to a draft Galápagos giant tortoise reference genome, as suggested by others (Shafer *et al.* 2016). However, our ddRAD-seq SNP data were called without mapping to a reference, so this methodological difference cannot completely explain the loss of signal.

Conclusions

421 Reduced-representation sequencing offers practical ways to take advantage of the
422 power of population genomics, even when samples and funds are limited (Narum *et al.* 2013).
423 Yet, thoughtful study design remains an essential component. Our analyses clearly showed that
424 tortoises representing each of three named species exhibit high genetic differentiation at the
425 genomic level, as demonstrated through high and significant F_{ST}, and positive F_{IT} estimates, as
426 well as through principal component and Bayesian clustering analyses. Using thousands of
427 SNPs gives high statistical power to detect population structure even when sample sizes of
428 individuals are as few as two or three individuals. However, the heterogeneity of samples within
429 a population can confound calculations using small sample sizes in unpredictable ways.
430 Reduced sample size also limits the diversity of analyses that can be performed, especially
431 limiting those that do not rely on a priori population designation, such as principal component
432 analysis and Bayesian clustering algorithms. Ultimately, we found that both our ddRAD-seq data
433 and a reanalysis of RNA-seq data generated by Loire *et al.* (2013) were consistent with the
434 findings of earlier microsatellite and mtDNA studies. We therefore expect genome-wide SNPs to
435 support the conclusions of population genetic studies of Galápagos giant tortoises beyond the
436 three species considered here.

437 Distinguishing populations and evolutionary lineages, such as the giant tortoise species
438 analyzed here, is a vital role for population genetic analyses to play in conservation (Funk *et al.*
439 2012). Results from such analyses can assist in protected area designation (Larson *et al.* 2014),
440 inform appropriate legal protections (vonHoldt et al. 2016a), and guide captive breeding
441 strategies (de Cara *et al.* 2011; Lew *et al.* 2015). We show that, as long as population genetics
442 theory is carefully taken into account, the use of genome-wide data enabled by high-throughput
443 sequencing can be a powerful tool in these conservation efforts, even when sample sizes are
444 limited.

445

446 **Data Accessibility:** Raw data from Illumina sequencing will be deposited to the NCBI Short
447 Read Archive (SRA) for all individuals included in this study. The vcf file used in the analyses
448 will be deposited on Dryad. Microsatellite genotypes and mitochondrial DNA sequences used in
449 the supplementary material are available upon request.

450

451 **Literature Cited:**

452 Allendorf FW (2017) Genetics and the conservation of natural populations: allozymes to
453 genomes. *Molecular Ecology*, **26**, 420–430.
454 Allendorf FW, Luikart G (2009) *Conservation and the genetics of populations*. Blackwell
455 Publishing Inc.
456 Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of
457 RADseq for ecological and evolutionary genomics. *Nat Rev Genet*, **17**, 81–92.
458 Beheregaray LB, Ciofi C, Caccone A, Gibbs JP, Powell JR (2003a) Genetic divergence,
459 phylogeography and conservation units of giant tortoises from Santa Cruz and Pinzon,
460 Galapagos Islands. *Conservation Genetics*, **4**, 31–46.
461 Beheregaray LB, Ciofi C, Geist D *et al.* (2003b) Genes record a prehistoric volcano eruption in
462 the Galapagos. *Science*, **302**, 75.
463 Beheregaray LB, Gibbs JP, Havill N *et al.* (2004) Giant tortoises are not so slow: Rapid
464 diversification and biogeographic consensus in the Galapagos. *Proceedings of the National*
465 *Academy of Sciences of the United States of America*, **101**, 6514–6519.
466 Brauer CJ, Hammer MP, Beheregaray LB (2016) Riverscape genomics of a threatened fish
467 across a hydroclimatically heterogeneous river basin. *Molecular Ecology*, **25**, 5093–5113.
468 Caccone A, Gentile G, Gibbs JP *et al.* (2002) Phylogeography and history of giant Galapagos
469 tortoises. *Evolution*, **56**, 2052–2066.
470 Caccone A, Gibbs JP, Ketmaier V, Suatoni E, Powell JR (1999) Origin and evolutionary
471 relationships of giant Galapagos tortoises. *Proceedings of the National Academy of*
472 *Sciences of the United States of America*, **96**, 13223–13228.
473 de Cara MÁR, Fernández J, Toro MA, Villanueva B (2011) Using genome-wide information to
474 minimize the loss of diversity in conservation programmes. *Journal of Animal Breeding and*
475 *Genetics*, **128**, 456–464.
476 Chang C, Chow C, Tellier L *et al.* (2015) Second-generation PLINK: rising to the challenge of
477 larger and richer datasets. *GigaScience*, **4**.
478 Chaves JA, Cooper EA, Hendry AP *et al.* (2016) Genomic variation at the tips of the adaptive
479 radiation of Darwin's finches. *Molecular Ecology*, **25**, 5282–5295.
480 Ciofi C, Milinkovitch MC, Gibbs JP, Caccone A, Powell JR (2002) Microsatellite analysis of
481 genetic divergence among populations of giant Galapagos tortoises. *Molecular Ecology*, **11**,
482 2265–2283.
483 Ciofi C, Wilson GA, Beheregaray LB *et al.* (2006) Phylogeographic history and gene flow among
484 giant Galapagos tortoises on southern Isabela Island. *Genetics*, **172**, 1727–1744.
485 Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools.
486 *Bioinformatics*, **27**, 2156–2158.
487 Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
488 *Bioinformatics*, **30**, 1844–1849.
489 Edwards DL, Benavides E, Garrick RC *et al.* (2013) The genetic legacy of Lonesome George

490 survives: Giant tortoises with Pinta Island ancestry identified in Galapagos. *Biological*
491 *Conservation*, **157**, 225–228.
492 Edwards DL, Garrick RC, Tapia W, Caccone A (2014) Cryptic structure and niche divergence
493 within threatened Galápagos giant tortoises from southern Isabela Island. *Conservation*
494 *Genetics*, **15**, 1357–1369.
495 Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using
496 high-throughput sequencing. *Proceedings of the National Academy of Sciences of the*
497 *United States of America*, **107**, 16196–16200.
498 Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using
499 the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.
500 Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus
501 genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.
502 Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating
503 conservation units. *TREE*, **27**, 489–496.
504 Garrick RC, Benavides E, Russello MA *et al.* (2012) Genetic rediscovery of an "extinct"
505 Galapagos giant tortoise species. *Current Biology*, **22**, R10–R11.
506 Garrick RC, Benavides E, Russello MA *et al.* (2014) Lineage fusion in Galapagos giant tortoises.
507 *Molecular Ecology*, **23**, 5276–5290.
508 Garrick RC, Kajdacsi B, Russello MA *et al.* (2015) Naturally rare versus newly rare:
509 demographic inferences on two timescales inform conservation of Galápagos giant
510 tortoises. *Ecology and Evolution*, **5**, 676–694.
511 Henry P, Russello MA (2013) Adaptive divergence along environmental gradients in a climate-
512 change-sensitive mammal. *Ecology and Evolution*, **3**, 3906–3917.
513 Keenan K, McGinnity P, Cross TF, Crozier WW, Prodöhl PA (2013) diveRsity: An R package for
514 the estimation and exploration of population genetics parameters and their associated
515 errors. *Methods in Ecology and Evolution*, **4**, 782–788.
516 Kimura M, Crow JF (1963) The measurement of effective population number. *Evolution*, **17**,
517 279–288.
518 Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program
519 for identifying clustering modes and packaging population structure inferences across K.
520 *Molecular Ecology Resources*, **15**, 1179–1191.
521 Larson WA, Seeb LW, Everett M V *et al.* (2014) Genotyping by sequencing resolves shallow
522 population structure to inform conservation of Chinook salmon (Oncorhynchus
523 tshawytscha). *Evolutionary Applications*, **7**, 355–369.
524 Lew RM, Finger AJ, Baerwald MR *et al.* (2015) Using Next-Generation Sequencing to Assist a
525 Conservation Hatchery: a Single-Nucleotide Polymorphism Panel for the Genetic
526 Management of Endangered Delta Smelt. *Transactions of the American Fisheries Society*,
527 **144**, 767–779.
528 Li H, Durbin R (2011) Inference of human population history from individual whole-genome
529 sequences. *Nature*, **475**, 493–496.
530 Loire E, Chiari Y, Bernard A *et al.* (2013) Population genomics of the endangered giant
531 Galapagos tortoise. *Genome Biology*, **14**.
532 Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them.
533 *Molecular Ecology*, **24**, 3223–3231.
534 Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-
535 sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.
536 Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of
537 individuals. *Genetics*, **89**, 583–590.
538 Payseur BA, Jing P (2009) A Genomewide Comparison of Population Structure at STRPs and

539 Nearby SNPs in Humans. *Molecular Biology and Evolution*, **26**, 1369–1377.
540 Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An
541 Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model
542 Species. *PLOS ONE*, **7**, e37135.
543 Poulakakis N, Edwards DL, Chiari Y *et al.* (2015) Description of a New Galapagos Giant
544 Tortoise Species (Chelonoidis; Testudines: Testudinidae) from Cerro Fatal on Santa Cruz
545 Island. *PLoS ONE*, **10**, e0138779.
546 Poulakakis N, Glaberman S, Russello M *et al.* (2008) Historical DNA analysis reveals living
547 descendants of an extinct species of Galapagos tortoise. *Proceedings of the National*
548 *Academy of Sciences of the United States of America*, **105**, 15464–15469.
549 Poulakakis N, Russello M, Geist D, Caccone A (2012) Unravelling the peculiarities of island life:
550 vicariance, dispersal and the diversification of the extinct and extant giant Galapagos
551 tortoises. *Molecular Ecology*, **21**, 160–173.
552 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus
553 genotype data. *Genetics*, **155**, 945–959.
554 Putman AI, Carbone I (2014) Challenges in analysis and interpretation of microsatellite data for
555 population genetic studies. *Ecology and Evolution*, **4**, 4399–4428.
556 Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population
557 history. *Nature*, **461**, 489–494.
558 Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to
559 environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–
560 4370.
561 Robinson JA, Ortega-Del Vecchyo D, Fan Z *et al.* (2016) Genomic flatlining in the endangered
562 island fox. *Current Biology*, **26**, 1183–1189.
563 Russello MA, Beheregaray LB, Gibbs JP *et al.* (2007a) Lonesome George is not alone among
564 Galápagos tortoises. *Current Biology*, **17**, R317–R318.
565 Russello MA, Glaberman S, Gibbs JP *et al.* (2005) A cryptic taxon of Galapagos tortoise in
566 conservation peril. *Biology Letters*, **1**, 287–290.
567 Russello MA, Hyseni C, Gibbs JP *et al.* (2007b) Lineage identification of Galapagos tortoises in
568 captivity worldwide. *Animal Conservation*, **10**, 304–311.
569 Shafer ABA, Peart CR, Tusso S *et al.* (2016) Bioinformatic processing of RAD-seq data
570 dramatically impacts downstream population genetic inference. *Methods in Ecology and*
571 *Evolution*, n/a-n/a.
572 Sork VL, Squire K, Gugger PF *et al.* (2016) Landscape genomic analysis of candidate genes for
573 climate adaptation in a California endemic oak, Quercus lobata. *American Journal of*
574 *Botany* , **103**, 33–46.
575 vonHoldt BM, Cahill JA, Fan Z *et al.* (2016a) Whole-genome sequence analysis shows that two
576 endemic species of North American wolf are admixtures of the coyote and gray wolf.
577 *Science Advances*, **2**.
578 vonHoldt BM, Kays R, Pollinger JP, Wayne RK (2016b) Admixture mapping identifies
579 introgressed genomic regions in North American canids. *Molecular Ecology*, **25**, 2443–
580 2453.
581 Wahlund S (1928) Zusammensetzung von populationen und korrelationserscheinungen vom
582 standpunkt der vererbungslehre aus betrachtet. *Hereditas*, **11**, 65–106.
583 Weir BS, Cocker (1984) Estimating F-Statistics for the Analysis of Population Structure.
584 *Evolution*, **38**, 1358–1370.
585 Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of Genetic Differentiation Measured
586 by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers.
587 *PLoS ONE*, **7**, e42649.

588 Wright B, Morris K, Grueber CE *et al.* (2015) Development of a SNP-based assay for measuring
589 genetic diversity in the Tasmanian devil insurance population. *BMC Genomics*, **16**.
590
591

592

593 **Table 1.** Number of polymorphic loci present in all individuals (*n*=10 per species) used for
594 analyses of each population (diagonal) and population pair (below diagonal).

	PBL (*C. becki*)	CRU (*C. porteri*)	VA (*C. vandenburghi*)
PBL (*C. becki*)	9,580		
CRU (*C. porteri*)	19,654	11,703	
VA (*C. vandenburghi*)	13,520	16,432	5,732

595
596

597 **Table 2.** Pairwise F_{ST} values between given species pairs. Above the diagonal, values
598 calculated using our dataset of SNPs with no missing data and common to the population pair,
599 along with 95% confidence intervals. Below the diagonal, values calculated using 12
600 microsatellite loci from Garrick et al. (2015) (see supplementary material section VIII). Data
601 were obtained using 10 samples for each population (PBL, VA, CRU) for the three species.

	PBL (*C. becki*)	CRU (*C. porteri*)	VA (*C. vandenburghi*)
PBL (*C. becki*)	xx	0.169 *(0.164 – 0.174)*	0.181 *(0.175 – 0.187)*
CRU (*C. porteri*)	0.137	xx	0.233 *(0.226 – 0.240)*
VA (*C. vandenburghi*)	0.163	0.202	xx

602
603

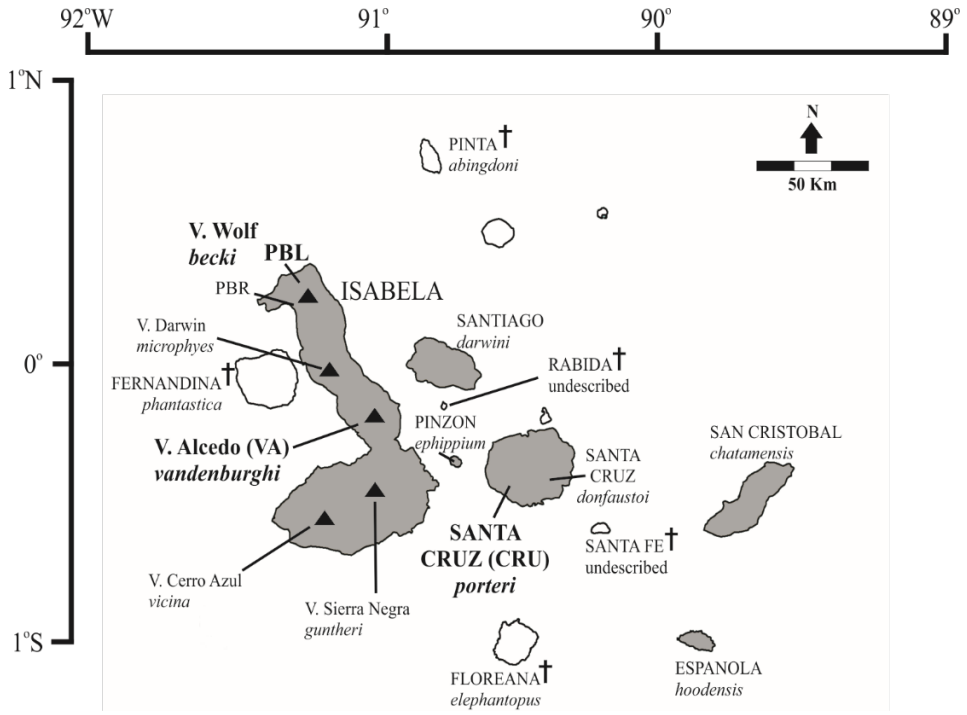


Figure 1.

615 **Figure 1.** Distribution map of Galápagos giant tortoises throughout the archipelago. The
616 islands with extant species are shown in gray, while the islands with extinct species are
617 in white. Black triangles identify the location of the four volcanoes on Isabela Island,
618 each with its own locally endemic tortoise species. Extinct species are identified by a
619 cross symbol. Names of each species are in cursive with a black line pointing to the
620 island or location within an island where they occur. The populations from the three
621 species in this study are identified by two or three letter symbols in bold: CRU = *C.*
622 *porteri*, Santa Cruz Island (La Caseta). VA = *C. vandenburghi*, Volcano Alcedo, central
623 Isabela Island, and PBL = *C. becki*, Piedras Blancas, Volcano Wolf, northern Isabela
624 Island.
625
626
627
628
629
630
631

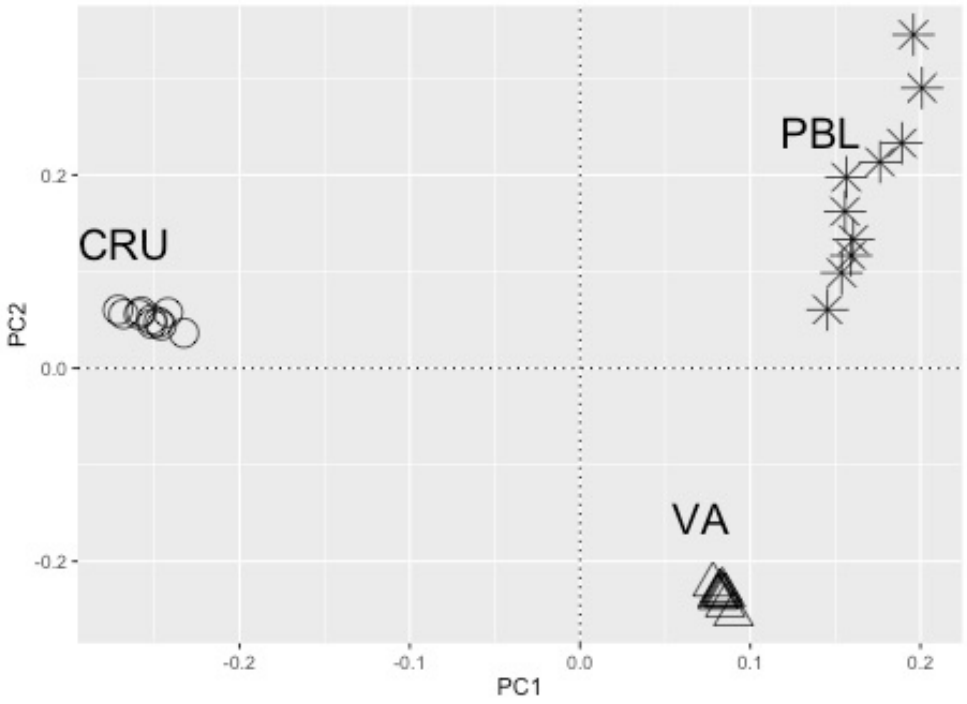
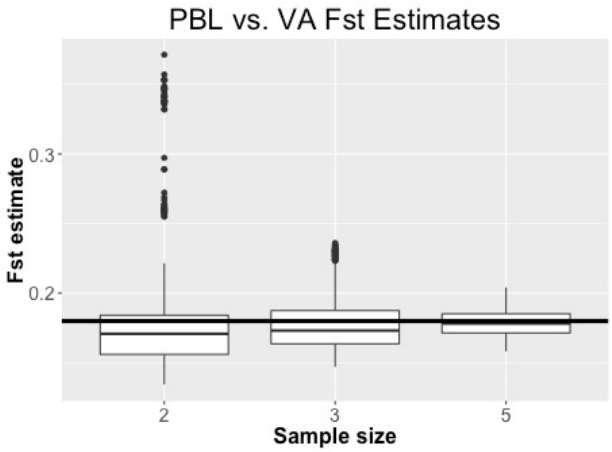


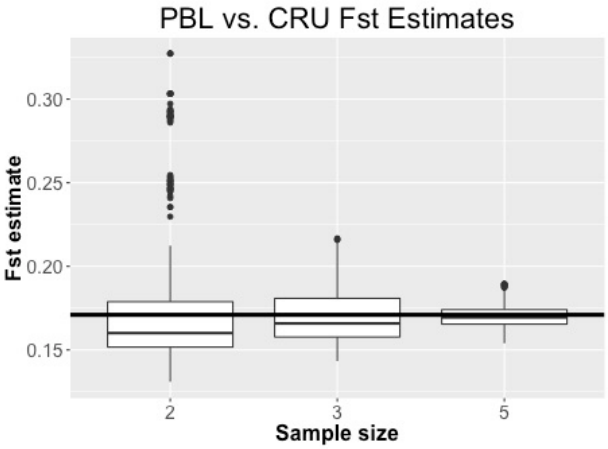
Figure 2.

646 **Figure 2**. Principal component 1 (PC1) plotted against principal component 2 (PC2) for 30
647 individuals from 3 populations, resulting from PCA analysis on 23,057 SNPs. Stars, open
648 circles, and open triangles identify individuals from the PBL (*C. becki*), CRU (*C. porteri*),
649 and VA (*C. vandenburghi*) populations, respectively. The analysis was carried out using
650 PLINK (Chang *et al.* 2015).
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676

677 A)

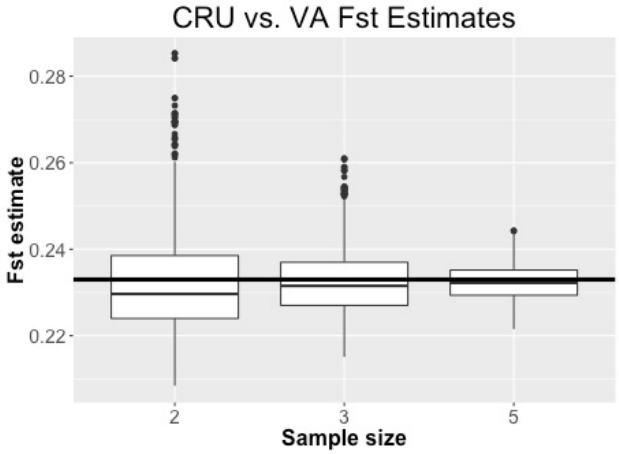


B)



678

679 C)



680

681 **Figure 3**

682

683

684

685

686

687

688

689

690 **Figure 3.** Boxplots of pairwise F_{ST} estimates using 1,000 randomly drawn subsamples of
691 individuals for each sample size (n=2, 3, or 5) from each population. PBL, CRU and VA
692 correspond to population samples from *C. becki*, *C. porteri* and *C. vandenburghi*,
693 respectively. The horizontal black line in each boxplot marks the F_{ST} value calculated
694 using all ten individuals from each population in the pairwise comparison (see
695 supplementary table S1). Lower hinge corresponds to first quartile (25th percentile);
696 upper hinge corresponds to third quartile (75th percentile). Whiskers indicate points
697 within 1.5 times the interquartile range (IQR), with outliers indicated as points beyond
698 that range.
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714